Event-triggered reinforcement learning; an application to buildings' micro-climate control

Ashkan Haji Hosseinloo[1], **Munther Dahleh**[2]
Laboratory for Information and Decision Systems, MIT, USA
ashkanhh@mit.edu[1], dahleh@mit.edu[2]

Abstract

Smart buildings have great potential for shaping an energy-efficient, sustainable, and more economic future for our planet as buildings account for approximately 40% of the global energy consumption. However, most learning methods for micro-climate control in buildings are based on Markov Decision Processes with fixed transition times that suffer from high variance in the learning phase. Furthermore, ignoring its continuing-task nature the micro-climate control problem is often modeled and solved as an episodic-task problem with discounted rewards. This can result in a wrong optimization solution. To overcome these issues we propose an event-triggered learning control and formulate it based on Semi-Markov Decision Processes with variable transition times and in an average-reward setting. We show via simulation the efficacy of our approach in controlling the micro-climate of a single-zone building.

Introduction

Buildings account for approximately 40% of global energy consumption about half of which is used by heating, ventilation, and air conditioning (HVAC) systems, the primary means to control micro-climate in buildings. Furthermore, buildings are responsible for one-third of global energy-related greenhouse gas emissions. Hence, even an incremental improvement in the energy efficiency of buildings and HVAC systems goes a long way towards building a greener, more economic, and energy-efficient future. In addition to their economic and environmental impacts, HVAC systems can also affect productivity and decision-making performance of occupants in buildings via controlling indoor thermal and air quality. For all these reasons micro-climate control in buildings is an important issue for its large-scale economic, environmental, and societal effects.

The main goal of the micro-climate control in buildings is to minimize the building's (mainly HVAC's) energy consumption while improving occupants' comfort in some metric. Model-based control strategies are often inefficient in practice, due to the complexity in building thermal dynamics and heterogeneous environment disturbances (Wei, Wang, and Zhu 2017). They also rely on an accurate model of the building that makes them resource-extensive and costly. Moreover, the need for prior modeling of the buildings prevents a plug and play deployment of the model-based controllers. To remedy these issues data-driven approaches for HVAC control have attracted much interest in the recent years towards building *smart* homes. Although the idea of *smart* homes where household devices (e.g. appliances, thermostats, and lights) can operate efficiently in an autonomous, coordinated, and adaptive fashion, has been around for a couple of decades (Mozer 1998), its realization now looks ever more pragmatic with immense recent advances in Internet of Things (IoT) and sensor technology (Minoli, Sohraby, and Occhiogrosso 2017). Among different data-driven control approaches, reinforcement learning (RL) has found more attention in the recent years due to enormous recent algorithmic advances in this field as well as its ability to learn efficient control policies solely from experiential data via trial and error.

The Neural Network House project (Mozer 1998; Mozer and Miller 1997) is perhaps the first application of RL in building energy management system. Since then and over the past couple of decades different RL techniques from tabular Q-learning (Liu and Henze 2006; Barrett and Linder 2015; Cheng et al. 2016; Chen et al. 2018) to Deep RL (Wei, Wang, and Zhu 2017; Avendano et al. 2018) have been employed to optimally control the micro-climate in buildings. The control objective in all these studies is a variation of energy consumption/cost minimization subject to some constraints e.g. occupants' comfort in some metric. More recently policy gradient RL techniques were adopted for the HVAC control problem. For instance, Deep Deterministic Policy Gradient (DDPG) was used in (Gao, Li, and Wen 2019) and (Li et al. 2019) to control energy consumption in a single-zone laboratory and 2-zone data center buildings, respectively. The reader is referred to (Hosseinloo et al. 2020) for a comprehensive literature review on RL application in smart buildings.

Similar to many other RL application studies in physical sciences, there are two main issues with the above-mentioned studies; first, they model and solve the problem

of micro-climate control as an *episodic-task* problem with *discounted reward* while it should be modeled as a *continuing-task* problem with *average reward*. Average reward is really what matters in continuing-task problems and greedily maximizing discounted future value does not necessarily maximize the average reward (Naik et al. 2019). In particular, solutions that fundamentally rely on episodes are likely to fare worse than those that fully embrace the continuing task setting.

Second, in all these studies the control problem is modeled based on Markov Decision Processes (MDPs) where learning and decision making occur at fixed sampling rate. The fixed time intervals between decisions (control actions) is restrictive in continuous-time problems; a large interval (low sampling rate) deteriorates the control accuracy while a small interval (high sampling rate) could drastically affect the learning quality. For instance, as reported in (Munos 2006) among others, policy gradient estimate is subject to variance explosion when the discretization time-step tends to zero. The intuitive reason for that problem lies in the fact that the number of decisions before getting a meaningful reward grows to infinity. Furthermore, learning and control at fixed time intervals may not be desired in large-scale resource-constrained wireless embedded control systems (Heemels, Johansson, and Tabuada 2012).

In this study, we eliminate the major drawbacks of the learning techniques discussed above by proposing an event-triggered learning controller where the control problem is formulated based on Semi-Markov Decision Processes (SMDPs) with variable time intervals (decision epochs). The problem is formulated in an RL framework as a continuing-task problem with undiscounted average-reward optimization objective. The rest of the paper is organized as follows. The next section explains the problem statement and the proposed controller. SMDP formulation section describes the problem formulation and the proposed learning framework. Finally, the simulation results and the paper remarks are presented in the last two sections.

Problem statement

In this study we present and explain our proposed learning methods via a simplified one-zone building; however, the methods and concepts are applicable to more general settings. Here we study the problem of minimizing the energy consumption in a one-zone building with unknown thermal dynamics and subject to occupants' comfort constraints. For specificity and with no loss of generality we consider the heating problem rather than cooling. Temperature of the building evolves as:

$$\frac{dT}{dt} = f(T; T_o, u), \tag{1}$$

where, $T(t) \in \mathbb{R}$ represents the building temperature, $T_o(t) \in \mathbb{R}$ is the outside temperature (disturbance), and $u(t) \in \{0, 1\}$ denotes the heater's ON/OFF status (the actual control actions). Unknown and potentially nonlinear thermal dynamics of the system are characterized by the function $f(.)$. Via the control action $u(t)$ we would like to maximize the performance measure J defined as:

$$J = \lim_{T\to\infty} \frac{1}{T} \int_0^T \{r_e u(t) + r_c (T - T_d)^2 + r_{sw} \delta(t - t_{sw})\}\, dt, \tag{2}$$

where, t_{sw} is the time when the controller switches from 0 to 1 (the heater switches from OFF to ON) or vice versa, and $\delta(.)$ is the Dirac delta function. The first term of the integrand penalizes the energy consumption while the second and the third terms correspond to occupants' comfort. Specifically the second term penalizes temperature deviations from a desired set-point temperature (T_d) while the third term prevents frequent ON/OFF switching of the heater. The relative effects of these terms are balanced by their corresponding weights i.e. r_e, r_c, and r_{sw}.

To reduce the space of possible control policies(laws) we constraint the optimization to a class of parameterized control policies, specifically to threshold policies. This strategy is particularly beneficial in the RL framework since it could significantly reduce the learning sample complexity. We characterize the threshold policies by some manifolds in the state space of the system which determine when the control action switches (e.g. ON $\rightleftarrows$ OFF in this study). We call these manifolds *switching manifolds* and the control action switches only when hitting these manifolds which we refer to as *events*. Figure 1 (a) illustrates an schematic threshold policy for the 1-zone building example with switching ON and OFF manifolds while Fig.1(b) depicts the thermal dynamics of the building temperature under such controller. We can mathematically formulate the control action as:

$$u(t) = \begin{cases} 0, & \text{if } T(t) \geq T^{\theta}_{\text{OFF}} \\ 1, & \text{if } T(t) \leq T^{\theta}_{\text{ON}} \\ u(t^-), & \text{otherwise} \end{cases}, \tag{3}$$

where, T^{θ}_{OFF} and T^{θ}_{ON} are thresholds (manifolds) for switching OFF and ON, respectively that are parameterized by parameter vector θ. These thresholds are in general state-dependent. The goal is to find the optimal control policy $u^*(t)$ within the parameterized policies i.e. to find the optimal parameter vector θ^*, which maximizes the long-run average reward (performance metric) J defined by (2) with no prior knowledge of the system dynamics. In the next section we cast this decision making problem as an SMDP.

SMDP formulation

By defining the switching manifolds the control problem is reduced to learning the optimal manifolds. Once the manifolds (the θ vector) are decided, the actual control actions ($u(t) \in \{0, 1\}$) are automatically known based on (3). We can thus think of the manifolds, hence the θ's as the higher level control actions and the ON/OFF heater status ($u(t)$) as the lower-level control actions. These are usually referred to as options and primitive actions in the hierarchical RL framework (Sutton, Precup, and Singh 1999). By doing so we are changing our decision variables

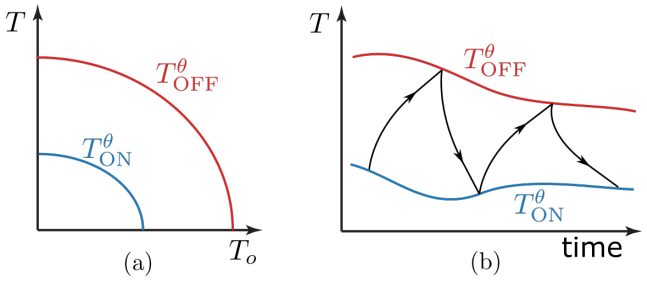

Figure 1: (a) Threshold policy and the switching manifolds for the 1-zone building example (b) thermal dynamics of the building example under threshold control policy

from $u(t)$ to θ. Although we could control (i.e. set the θ values) and learn (i.e. learn a better θ) at fixed time steps we restrict them to times when the events occur i.e. when the system state trajectory hits a manifold. We do this because making too many decision in a short period of time (with no significant accumulated reward) could result in large variance as discussed earlier. This change in the timing of the control and learning changes the underlying formulation from an MDP with fixed transition times to an SMDP with stochastic transition times.

We study the control problem in an RL framework in which an agent acts in a stochastic environment/system by sequentially choosing actions with no knowledge of the environment/system dynamics. We model the RL control problem as an SMDP which is defined as a five tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, R, F)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}$ is a set of state-action dependent transition probabilities, R is the reward function, and F is a function giving probability of transition time, aka sojourn or dwell time, for each state-action pair. Let τ_k be the decision epochs (times) with $\tau_0 = 0$, and $S_k \in \mathcal{S}$ be the state variable at decision epoch τ_k. If the system is at state $S_k = s_k$ at epoch τ_k and action $A_k = a_k$ is applied, the system will move to the state $S_{k+1} = s_{k+1}$ at epoch τ_{k+1} with probability $p(s_{k+1}|s_k, a_k) = \mathcal{P}(S_{k+1} = s_{k+1}|S_k = s_k, A_k = a_k)$. This transition occurs within t_k unit times with a probability of $F(t_k|s_k, a_k) = \Pr(\tau_{k+1} - \tau_k \leq t_k|S_k = s_k, A_k = a_k)$. Hence, the SMDP kernel $\Phi(s_{k+1}, t_k) = \Pr(S_{k+1} = s_{k+1}, \tau_{k+1} - \tau_k \leq t_k|S_k = s_k, A_k = a_k)$ could be written as $\Phi(s_{k+1}, t_k) = p(s_{k+1}|s_k, a_k)F(t_k|s_k, a_k)$.

The reward function for an SMDP is in general more complex than that of an MDP. Between epochs ($\tau_k \leq t' \leq \tau_{k+1}$) the system evolves based on the so-called *natural process* $W_{t'}$. Let us suppose the reward between two decision epochs consists of two parts; a fixed state-action dependent reward of $f(s_k, a_k)$ and a time-continuous reward accumulated in the transition time at a rate of $c(W_{t'}, s_k, a_k)$. We can then write the expected reward $r(s_k, a_k) \in R(S_k, A_k)$ between two epochs of τ_k and τ_{k+1} as:

$$r(s_k, a_k) = f(s_k, a_k) + \mathbb{E}\left[\int_{\tau_k}^{\tau_{k+1}} c(W_{t'}, s_k, a_k)dt'|S_k = s_k, A_k = a_k\right]. \tag{4}$$

Let us also define the average transition time starting at state s_k and under action a_k as $\tau(s_k, a_k)$:

$$\begin{aligned}\tau(s_k, a_k) =& \mathbb{E}\left[\tau_{k+1} - \tau_k|S_k = s_k, A_k = a_k\right] \\ =& \int_0^{\infty} tF(dt|s_k, a_k).\end{aligned} \tag{5}$$

The actions a_ks of the SMDP are determined by a stochastic or deterministic policy in each state. In many real-world control problems the optimal and/or the desired control policy is a deterministic policy. Hence, here we focus on deterministic policies $a = \mu(s)$ which deterministically map the state s to the action a. Furthermore, as discussed earlier, for the sake of scalability and sample efficiency we restrict the control problem to a class of policies $\mu^{\theta}(s)$ parameterized by the parameter vector θ. With this assumption the expected rewards and the transition times at each state will be functions of the states and the parameter vector θ i.e. $r(s_k, a_k) = r(s_k; \theta)$ and $\tau(s_k, a_k) = \tau(s_k; \theta)$. Then the infinite-horizon average reward could be written as[1]:

$$J(\theta) = \lim_{n\to\infty} \frac{\mathbb{E}\left[\sum_{k=0}^{n} r(s_k; \theta)\right]}{\mathbb{E}\left[\sum_{k=0}^{n} \tau(s_k; \theta)\right]}. \tag{6}$$

An online learning algorithm could be devised if we can calculate a good estimate of the gradient $\nabla_{\theta} J$ in an online fashion, which could then be used to improve the policy parameters via stochastic gradient. But let us first draw clear connections between the SMDP formulation presented in this section and the micro-climate control problem in the previous section. By defining the switching manifolds temperature thresholds become the actions of the underlying SMDP. Let us take the building temperature (T) and the heater status (h) at the beginning of epochs as the state of the system i.e. $s_k = [T_k, h_k]$. Then we can write actions as $a = \mu^{\theta}(s) = h\,T^{\theta}_{\text{OFF}}(s) + (1 - h)T^{\theta}_{\text{ON}}(s)$, where $T^{\theta}_{\text{OFF}}(s)$ and $T^{\theta}_{\text{ON}}(s)$ are threshold temperatures for switching the heater `OFF` and `ON`, respectively and they could generally be state-dependent. Regarding the rewards, by comparing equations (2) and (4) one can conclude that $f(s_k, a_k) = r_{sw}$ and $c(W_{t'}, s_k, a_k) = r_e u(t') + r_c(T(t') - T_d)^2$.

If $r(s, a)$ and $\tau(s, a)$ are known and we somehow have access to the system dynamics, we can estimate $J(\theta)$ for a given parameter vector θ by constructing a long sequence of $s_0, a_0, r_0, \tau_0, ..., s_n, a_n, r_n, \tau_n$ via simulation. If we do this for different values of θ we can approximate the performance metric J as a function of θ. We can then use this approximation to estimate the performance gradient and use

[1]For the average reward to be independent of the initial state, the embedded MDP is required to be unichain.

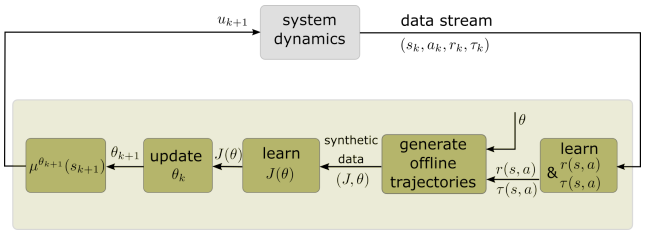

Figure 2: Block diagram of the online learning control

it to improve the actual policy via e.g. stochastic gradient. The idea here is to construct the above-mentioned trajectory sequence with online learning of $r(s,a)$ and $\tau(s,a)$ but without learning the system dynamics. This is possible because of our choice of policies, namely the threshold policies. Since the action a_k is a temperature threshold the temperature in the next epoch is automatically revealed at the current epoch i.e. $T_{k+1} = a_k$. Moreover, because these thresholds are switching manifolds the heater status must switch in the next epoch, i.e. $h_{k+1} = 1 - h_k$. However, in more complex set-ups, we may not be able to fully deduce the next state of the system via threshold policies. For instance, if the system state includes the electricity price, we cannot fully evaluate s_{k+1} based on s_k and a_k; but we can still construct a less-accurate sequence of transitions which could be sufficient since we usually do not need a very accurate estimate of $\nabla_\theta J$ for online learning. The online learning control explained here is schematically illustrated in the form of a block diagram in Fig. 2.

Results

In this section we implement our proposed method to control the heating system of a one-zone building in order to minimize energy consumption without jeopardizing the occupants' comfort. We use a simplified linear model characterized by a first-order ordinary differential equation as follows:

$$C\frac{dT}{dt} + K(T - T_o) = u(t)\dot{Q}_h, \tag{7}$$

where, $C = 2000\,kJK^{-1}$ is the building's heat capacity, $K = 325WK^{-1}$ is the building's thermal conductance, and $\dot{Q}_h = 13\,kW$ is the heater's power. As defined earlier, $h(t) \in \{0, 1\}$ is the heater status, and $T_o = -10\,^\circ C$ is the outdoor temperature. The reward rates are set as follows: $-r_{sw} = 0.8\,unit, -r_e = 1.2/3600\,unit\,s^{-1}$, and $r_c = -1.2/3600\,unit\,K^{-2}\,s^{-1}$.

The optimal control for this example is indeed a threshold policy with constant ON/OFF thresholds. Via brute-force simulations and search the optimal thresholds are found to be $T_{\text{ON}} = 12.5\,^\circ C$ and $T_{\text{OFF}} = 17.5\,^\circ C$ with a corresponding long-run average reward of $J = -3.70\,unit\,hr^{-1}$ (see Fig.3). This is the ground truth thresholds for the optimal control of the building which our learning controller (Fig. 2) should learn using an stream of online data.

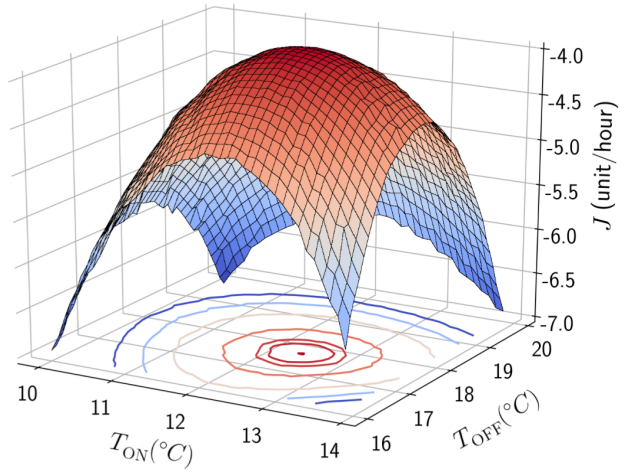

Figure 3: Long-run average reward (performance metric) J as a function of fixed ON/OFF temperature thresholds

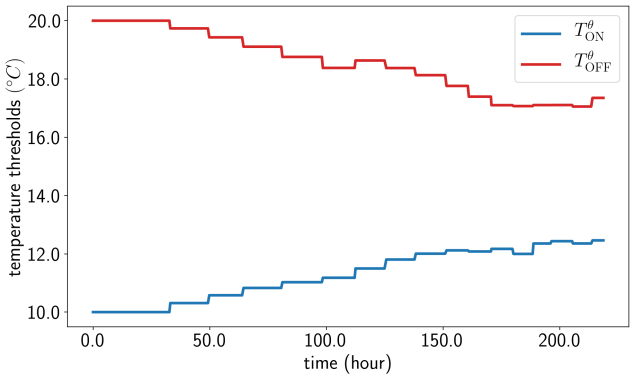

Figure 4: Time history of the learnt temperature thresholds during the online learning process

Since we know the optimal controller has fixed temperature thresholds, we can represent the control policy with only two parameters (2-component θ vector) i.e. the threshold themselves (T_{ON} and T_{OFF}). Also, we use neural nets for function approximation; a neural net with one hidden layer and 24 hidden nodes for $r(s,a)$ and $\tau(s,a)$, and another neural net with one hidden layer and 10 hidden nodes for $J(\theta)$. It is worth noting that our proposed control is an off-policy control. A very exploratory behaviour policy is employed for the learning simulation. Figure 4 illustrates how the controller learns the optimal thresholds within less than a week. The learnt thresholds at the end of the learning process are $T_{\text{ON}} = 12.5\,^\circ C$ and $T_{\text{OFF}} = 17.4\,^\circ C$ which are almost the same as the optimal thresholds.

Conclusion

In this study we proposed an SMDP framework for RL-based control of micro-climate in buildings. We utilized threshold policies in which the learning and control take place when the thresholds are reached. This results in

variable-time intervals for the learning and control which makes the SMDP framework more suitable for this class of control problems. Using the threshold policies we developed a model-based policy gradient RL approach for the controller. We showed via simulation the efficacy of our approach in controlling the micro-climate of a single-zone building.

References

Avendano, D. N.; Ruyssinck, J.; Vandekerckhove, S.; Van Hoecke, S.; and Deschrijver, D. 2018. Data-driven optimization of energy efficiency and comfort in an apartment. In *2018 International Conference on Intelligent Systems (IS)*, 174–182. IEEE.

Barrett, E., and Linder, S. 2015. Autonomous hvac control, a reinforcement learning approach. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 3–19. Springer.

Chen, Y.; Norford, L. K.; Samuelson, H. W.; and Malkawi, A. 2018. Optimal control of hvac and window systems for natural ventilation through reinforcement learning. *Energy and Buildings* 169:195–205.

Cheng, Z.; Zhao, Q.; Wang, F.; Jiang, Y.; Xia, L.; and Ding, J. 2016. Satisfaction based q-learning for integrated lighting and blind control. *Energy and Buildings* 127:43–55.

Gao, G.; Li, J.; and Wen, Y. 2019. Energy-efficient thermal comfort control in smart buildings via deep reinforcement learning. *arXiv preprint arXiv:1901.04693*.

Heemels, W.; Johansson, K. H.; and Tabuada, P. 2012. An introduction to event-triggered and self-triggered control. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, 3270–3285. IEEE.

Hosseinloo, A. H.; Ryzhov, A.; Bischi, A.; Ouerdane, H.; Turitsyn, K.; and Dahleh, M. A. 2020. Data-driven control of micro-climate in buildings; an event-triggered reinforcement learning approach. *arXiv preprint arXiv:2001.10505*.

Li, Y.; Wen, Y.; Tao, D.; and Guan, K. 2019. Transforming cooling optimization for green data center via deep reinforcement learning. *IEEE transactions on cybernetics*.

Liu, S., and Henze, G. P. 2006. Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory: Part 1. theoretical foundation. *Energy and Buildings* 38(2):142 – 147.

Minoli, D.; Sohraby, K.; and Occhiogrosso, B. 2017. Iot considerations, requirements, and architectures for smart buildings—energy optimization and next-generation building management systems. *IEEE Internet of Things Journal* 4(1):269–283.

Mozer, M. C., and Miller, D. 1997. Parsing the stream of time: The value of event-based segmentation in a complex real-world control problem. In *International School on Neural Networks, Initiated by IIASS and EMFCSC*, 370–388. Springer.

Mozer, M. C. 1998. The neural network house: An environment hat adapts to its inhabitants. In *Proc. AAAI Spring Symp. Intelligent Environments*, volume 58.

Munos, R. 2006. Policy gradient in continuous time. *Journal of Machine Learning Research* 7(May):771–791.

Naik, A.; Shariff, R.; Yasui, N.; and Sutton, R. S. 2019. Discounted reinforcement learning is not an optimization problem. *arXiv preprint arXiv:1910.02140*.

Sutton, R. S.; Precup, D.; and Singh, S. 1999. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence* 112(1-2):181–211.

Wei, T.; Wang, Y.; and Zhu, Q. 2017. Deep reinforcement learning for building hvac control. In *Proceedings of the 54th Annual Design Automation Conference 2017*, 22. ACM.